Choosing among models

September 18, 2014

Straightforward Bayes approach

- Treat "model number" M as a parameter to be estimated.
- The parameter is discrete, but this poses no problem in principle.
- Bayes rule is specified in terms of densities, but they can be densities w.r.t. measures with discrete components.

Example

- E.g. model 1 is $y \sim N(\mu, 1)$, model 2 is $y \sim N(0, \sigma^2)$. The parameter space is the real line (for μ), a separate copy of the positive real line (for σ^2), with model number telling us which of the two continuous spaces we are in.

- The natural base measure is Lebesgue measure on the two continuous components, together with discrete measure — unit weight on $M = 1$ and $M = 2$.

- A prior that puts equal weight on each model, a $N(0, 1)$ prior on μ, and a $\mathrm{Gamma}(1, 1)$ (exponential) prior on σ^2 would have a density function $.5\phi(\mu) + .5\exp(-\sigma^2$.

Marginal on M

- Then the posterior probabilities on the models are just the marginal distribution over M, with other parameters integrated out.

- In other words, the posterior probabilities are proportional to the prior probabilities π_i times the integral over the parameter space of prior density times likelihood for each model.

- If model i has pdf for the data $p(y, \theta_i, i)$, the ratio

$$\frac{\int p(y, \theta_i, i) q(\theta_i, i)\, d\theta_i}{\int p(y, \theta_j, j) q(\theta_j, j)\, d\theta_j}$$

where $q(\_, i)$ is the prior density for model i, is the **Bayes factor** for model i vs model j.

- The ratio of prior probabilities times the Bayes factor is the ratio of posterior probabilities.

Asymptotics

In large samples, likelihoods for nicely behaved models concentrate near a point and become approximately Gaussian in shape (i.e., their logs are approximately quadratic). If the likelihood for model i is nearly Gaussian, with $\hat{\theta}_i$ the MLE and

$$\Sigma_i = -\left(\frac{\partial^2 \log(p(y, \theta_i, i))}{\partial \theta_i \partial \theta_i'}\right)^{-1},$$

then the integrated likelihood should be approximately

$$p(y, \hat{\theta}_i, i)\, |\Sigma_i|^{.5}\, (2\pi)^{k/2}\,.$$

BIC

- Generally Σ_i behaves in large samples like $T^{-1}\bar{\Sigma}_i$. (It's $\Sigma^2(X'X)^{-1}$ for the SNLM.)

- So the log of this integrated likelihood is approximately

$$\log(p(y, \hat{\theta}_i, i)) - \frac{k}{2}\log T + \frac{1}{2}\log\left|\bar{\Sigma}_i\right| + \frac{k}{2}\log(2\pi)\ .$$

BIC2

- The last term doesn't change with T. In nicely behaved models, if the models are both "true", while one has more free parameters than the other, the $k \log T$ term dominates eventually, favoring the model with lower k.

- If one model is true, the other not, then the differences in the $\log(p(y, \hat{\theta}_i, i))$ term dominate, favoring the true model.

- BIC: compare difference in log likelihoods to $(k/2) \log T$

Model selection vs. model averaging

- A Bayesian approach does not suggest picking one model, unless posterior odds in favor of one model are extreme or there is a "cost" to dealing with more than one model.

- For prediction, a pure Bayesian approach would average predictions from all available models, weighting the predictions by the posterior probabilities on the models.

Issues with posteriors across models

- They depend on priors on the models. (Of course, but this is unlike the case of posteriors over a continuous parameter space, where smooth priors stop mattering in large samples.)

- They depend on the smooth priors within models.

- Within a continuous-parameter model, the likelihood concentrates in large samples in a region where the prior density is nearly constant. Normalization then makes the prior disappeaar.

- But in model comparison, the level of the nearly flat prior density affects odds ratios, even asymptotically.

How to manipulate posteriors on models

- If you give the parameters of a model a "nearly flat" prior, say a normal with very large variance, you can drive the posterior odds on that model to zero by increasing the variance of the prior.

- If you peek at the data and then specify for model j a "prior" that centers on the MLE, with very low variance, you can eliminate all the implicit penalties on large models or on models with high uncertainty about parameters.

- If you use as a prior what is actually a posterior for the same model and the same data, you will increase the posterior odds. Repeating this enough times gives you the same effect as a prior concentrated tightly on the MLE. People don't do often do this exactly, but if they use results of previous research with nearly the same data and models to arrive at a "prior", they may inadvertently approximate this.

Posterior odds on sparse sets of models

- In practice, posterior odds often apparently imply very high probabilities (e.g., .999) that one model among those being analyzed is the true model. This often seems an implausible degree of certainty.

- The fact that the result seems implausible suggests something wrong with the prior or the model. In this case, it is that the "model" starts out saying that one of a finite collection of models is correct. Usually we don't really quite believe this.

- Usually one can think of ways to expand or vary the model with implausibly high posterior probability that would expand the set of models to include others close to the high-probability one. In the expanded model space, the odds would no longer be so extreme.

Examples

- Model 1 includes only x, model 2 includes only y. Expand by including both. Or, model 1 includes x, model 2 does not and has high posterior probability. Expand by including a version of model 1 with small prior variance around zero on the coefficient of x.

- BDA's skepticism about posterior odds on models is based on believing that usually something like the first case is possible — embed your separate models in a larger model with continuous parameter space.